ВЕСТНИК ЮГОРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2022 г. Выпуск 4 (67). С. 120–132

УДК 681.3.06,681.322
DOI: 10.18822/byusu202204120-132

ПРИМЕНЕНИЕ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ ДЛЯ ОПТИМИЗАЦИИ РЕШЕНИЯ ЗАДАЧ ФИЛЬТРАЦИИ И ПРОГНОЗИРОВАНИЯ В ДИНАМИЧЕСКИХ СИСТЕМАХ ТЕСТИРОВАНИЯ ПРОГРАММ

Полухин Павел Валерьевич
кандидат технических наук,
преподаватель факультета прикладной математики,
информатики и механики,
ФГБОУ ВО «Воронежский государственный университет»
Воронеж, Россия
E-mail: alfa_force@bk.ru

Предмет исследования: вероятностные временные модели тестирования, предназначенные для формирования сложных стохастических связей между отдельными элементами тестирования и обнаружения определенных групп программных ошибок веб-приложений.

Цель исследования: обоснование возможности применения генетических алгоритмов в процессе решения вероятностных задач тестирования на основе многочастичного фильтра и оценка их эффективности. В исследовании приведены основополагающие методы, позволяющие повысить точность апостериорного распределения вероятностных моделей тестирования и общее число выборок, согласованных со свидетельствами.

Методы и объекты исследования: объектом исследования является решение задач фильтрации и сглаживания для вероятностной модели тестирования на основе многочастичного фильтра. Приведены методы и алгоритмы на основе метода Монте-Карло, позволяющие в сочетании с генетическими алгоритмами повысить точность получения апостериорных оценок выборок. Такой подход позволяет сузить область разброса выборок, а также увеличить степень их согласованности. Формирование каждой следующей выборки будет осуществляться с учетом предыдущих за счет применения операций скрещивания и мутации.

Основные результаты исследования: в результате доказана состоятельность предложенных подходов к решению задач фильтрации и прогнозирования в процессе реализации процедур тестирования на основе алгоритмов многочастичного фильтра и генетических алгоритмов. Приведенные практические результаты доказывают конструктивность и научную обоснованность предложенных методов и алгоритмов для решения задач тестирования веб-приложений.

Ключевые слова: многочастичный фильтр, метод Монте-Карло, цепь Маркова, скрытая марковская модель, генетический алгоритм.

APPLICATION OF GENETIC ALGORITHMS TO OPTIMIZE SOLUTION OF FILTERING AND PREDICTION PROBLEMS IN DYNAMIC PROGRAM TESTING SYSTEMS

Pavel V. Polukhin
Candidate of Technical Sciences,
Lecturer of the Faculty of Applied Mathematics, Informatics and Mechanics,
Voronezh State University
Voronezh, Russia
E-mail: alfa_force@bk.ru

Subject of research: probabilistic time models of testing created to form complex stochastic connections between individual test elements and developed to detect certain groups of the web applications program errors.

The purpose of research: substantiate the possibility of using genetic algorithms in the process of solving probabilistic testing problems based on a multi-particle filter and evaluate their effectiveness. The study provides fundamental methods to improve the accuracy of the posterior distribution of probabilistic testing models and the total number of matched with evidences samples.

Methods and objects of research: object of the research is to solve filtering and smoothing problems for a probabilistic test model based on a multi-particle filter. Methods and algorithms based on the Monte Carlo method are presented, allowing, in combination with genetic algorithms, to increase the accuracy of obtaining posterior estimates of samples. This approach allows you to narrow the range of samples, as well as increase their consistency. The formation of each next sample will be carried out taking into account the previous ones through the use of crossover and mutation operations.

The main results of research: as a result, the validity of the proposed approaches to solving filtration and prediction problems in the process of implementing testing procedures based on multi-particle filter algorithms and genetic algorithms was proved. The given practical results prove the constructiveness and scientific validity of the proposed methods and algorithms for solving web applications testing problems.

Keywords: particle filter, Monte-Carlo method, Markov chain, hidden Markov model, genetic algorithm

Введение

Генетические алгоритмы представляют собой универсальный инструмент, основанный на представлении естественного процесса эволюции. Применение генетических алгоритмов к математическим моделям было впервые рассмотрено в работах Холланда В его работах рассматривается применение теории естественного отбора Дарвина для решения ряда прикладных математических задач поиска, распознавания образов и статистического вывода. Согласно его предположению каждой модели, может быть противопоставлено ограниченное число структур α, представляющих собой некоторое множество хромосом (комбинация генов), изменение которых производит за счет мутации. В таком случае, общая сложность модели будет определяться общим числом генотипов (структура генов). Генетические алгоритмы с научно-практической точки зрения представляют особый интерес для оптимизации процедур обучения и логического вывода вероятностных моделей. Одним из основных алгоритмов решения данных задач является семейство алгоритмов на основе многочастичного фильтра. Решение задачи нелинейной фильтрации на основе генетических алгоритмов было рассмотрено в исследованиях Морала [1,2]. Суть предложенного подхода заключается в случайной мутации выборок, формируемых в соответствии с переходными вероятностями марковской цепи. В случае динамических вероятностных моделей данный подход не может быть явно применен, так как выборки, подвергнутые мутации предварительно должны соответствовать вероятностному распределению по всем свидетельствам. В связи с этим возникает необходимость пошаговой фильтрации с последующем применением генетического алгоритма уже к сформированному весовому распределению, полученному в соответствии с распределениями переходов и свидетельств для момента времени $t+k$. Основная особенность предлагаемого подхода заключается в возможности скрещивания выборок, исключительно имеющих общие свидетельства, следовательно, переменные, не имеющие условные связи, будут исключены из генетического алгоритма. В таком случае можно снизить область разброса выборок и повысить степень их согласования со свидетельствами. Для динамических моделей также допустимо использовать свидетельства со всех предыдущих состояний. Следовательно, можно прийти к выводу, что генетические алгоритмы также могут быть исполь-

зованы в процессе ретроспективного анализа таких моделей, при этом общее число генерируемых выборок для обеспечения требуемой точности может быть сокращено.

Результаты и обсуждение

Вероятностный вывод во временных моделях тестирования

Решение задач вероятностного вывода является одной из основных задач решаемых с помощью вероятностных моделей. Наибольший интерес представляют модели, состоящие из нескольких временных состояний. Среди таких моделей выделяют динамические байесовские (ДБС) сети и скрытые марковские модели (СММ). В исследовании рассмотрим СММ, так как в общем случае ДБС можно представить в виде СММ. СММ является графическим представлением марковского процесса, где каждому срезу соответствует модель перехода $\mathrm{P}\left(\mathrm{X}_1 \mid \mathrm{X}_2\right)$ и модель восприятия $\mathrm{P}\left(E_2 \mid \mathrm{X}_2\right)$. Каждому узлу СММ ставится в соответствие таблица условных вероятностей, характеризующая связи как в рамках одного среза, так и нескольких срезов. Типовая структура модели СММ приведена на рис. 1

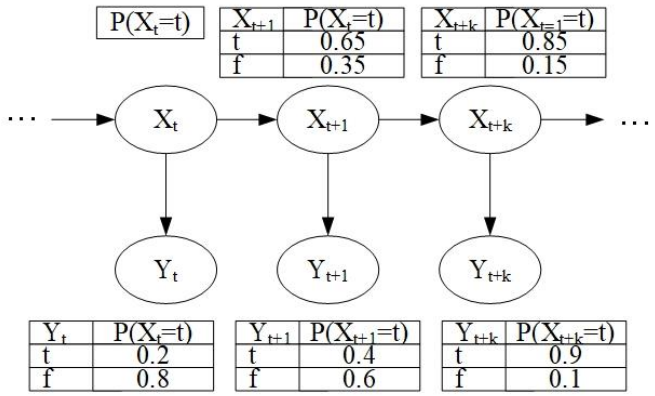

Рисунок 1 – Структура срытой марковской модели из трех временных срезов.

Матрица переходов, соответствующая рис. 1, будет иметь следующий вид:

$$T = P\left(X_{t+1} | X_t\right) = \begin{pmatrix} T_{11} & T_{12} & \cdots & T_{1k} \\ T_{21} & T_{22} & \cdots & T_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nk} \end{pmatrix}, \quad (1)$$

$$T_{ij} = P(X_{t+1} = j \mid X_t = i).$$

Тогда полное совместное распределение, соответствующее СММ можно записать в следующем виде [3,4]:

$$P\left(X, Y\right) = P\left(X_0\right) \prod_{t=1}^{k} P\left(X_{t+1} | X_t\right) \times$$

$$\times \prod_{j=1}^{k} P\left(X_{j+1} | X_{j+1}\right) = \quad (2)$$

$$= P\left(X_0\right) \prod_{t=1}^{k} T_{ij} \prod_{j=1}^{k} P\left(X_{j+1} | X_{j+1}\right),$$

где $P\left(X_0\right)$ – начальное распределение вероятностей, соответствующее моменту $\mathrm{t} = 0$.

Определив вероятностную семантику СММ, перейдем к рассмотрению алгоритмов вероятностного вывода. Среди различных подходов к решению задачи вероятностного вывода наибольший интересе представляет многочастичный фильтры (МЧФ). Сущность МЧФ заключается в формировании множества независимых друг от друга выборок с соответствующими им весами $\mathrm{W}(\mathrm{X},\mathrm{Y},\mathrm{E})$. Отметим, что $\mathrm{E} \subset Y$ – свидетельства, поступающие в виде непрерывного потока плоть до текущего состояния. В данном случае вес будет определять степень согласованности выборок S. В таком случае, условие согласования выборок S со свидетельством E запишем в следующем виде [4,6]:

$$\begin{aligned} & P\left(X_{t+1}|E_{1:t+1}\right)=\sum_{Y} N\left(X_{t+1},Y_{1:t},E_{1:t}\right)\times W \approx \\ & \approx \sum_{Y} S_{ws}\left(X_{t+1},Y_{1:t},E_{1:t}\right)W = \\ & =\sum_{Y} P\left(X_{t+1},Y_{1:t},E_{1:t}\right)=P\left(X_{t+1}|E_{1:t+1}\right), \\ & W=W\left(X_{t+1},Y_{1:t},E_{1:t}\right), \end{aligned} \tag{3}$$

где $N\left(X_{t+1},Y_{1:t},E_{1:t}\right)$ – число выборок, достигающих состояния t+1.

Существует несколько основных подходов к решению задач логического вывода на основе МЧФ. Наибольший интересе представляют МЧФ на основе выборки по значимости (ВЗ) и взвешивания на основе правдоподобия (ВП). В данном исследовании рассмотрим МЧФ с взвешиванием на основе правдоподобия. Данный алгоритм позволяет изначально производить генерацию только из области выборок $S(X_1, X_2, \dots, X_n)$, которые в наибольшей степени согласуются со свидетельствами $\mathrm{E} = (e_1, e_2, \dots, e_n)$. Такое условие достигается за счет того, что в процессе выполнения вероятностного вывода на основе ВП происходит определение и фиксирование переменных свидетельств E, а формирование выборок осуществляется исключительно для всех оставшихся переменных $Z = X \cup Y$, где X переменные запроса и Y переменные состояния. В процессе выполнения алгоритма производится развертывания динамической модели и происходит формирования выборок S_{ws}, которые взвешиваются с учетом их правдоподобия на основе анализа полученных весов $W\left(X_{t+1},Y_{1:t},E_{1:t}\right)$. Веса выборок рассчитываются на основе переходных вероятностей $P\left(X_{t+1}|X_t\right)$ и условного распределения по всем свидетельствам $P\left(E_{t+1}|X_{t+1}\right)$ Типовая схема фильтра МЧФ с ВП приведена на рис. 2

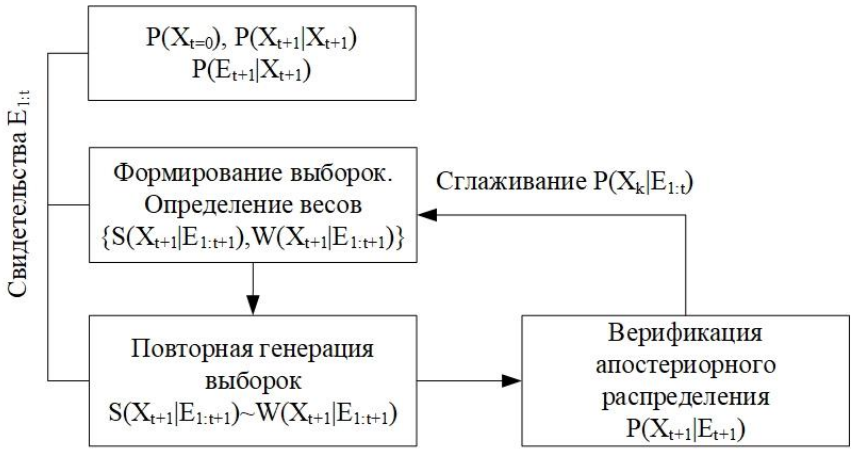

Рисунок 2 – Типовая схема МЧФ-фильтра с ВП

На рис. 2 $W(X_{t+1} \mid E_{1:t+1})$ – веса выборок S, $P(E_{t+1}|X_{t+1})$ и $P(X_{t+1}|X_t)$ модели перехода и восприятия. Рассматривая метод вероятностной оценки выборок на основе ВП, отметим главное его преимущество по сравнению с ВЗ. ВП исключает формирование приближенного распределения по значимости $Q(X_t \mid X_{0:t-1}, E_{1:t})$. Следовательно, нет необходимости на каждом этапе верификации апостериорного распределения $P(X_{t+1}|E_{t+1})$ производить вычисление дистанции Кульбака-Лейблера (КЛ) для оценки степени соответствия $Q(X_{t+1} \mid E_{t+1})$ и $P(X_{t+1}|E_{t+1})$. Вместо этого в ВП каждая выборка формируется в соответствии с весом $X_{t+1}^i \sim W(X_{t+1} \mid X_{1:t}, E_{1:t+1})$, при этом выборки с наименьшим весом исключаются из апостериорного распределения. Следовательно, согласно алгоритму ВП степень правдоподобия выборки будет определяться в соответствии с ее весом $W(X_{t+1} \mid X_{1:t}, E_{1:t+1})$. С учетом того, что распределение по всем выборкам прямо пропорционально общему числу выборок N, запишем распределение для выборок в соответствии с распределением $P(X_t \mid E_{1:t})$:

$$S(X_t \mid E_{1:t}) = N \times P(X_t \mid E_{1:t}), N \to \infty. \tag{4}$$

Используя переходное распределение вероятностей $P(X_{t+1}|X_t)$ для модели СММ, приведенной на рисунке 1, можно определить условное распределение по всем выборкам, достигшим состояния $t+1$ у четом распределения (4). Тогда получим следующую сумму по всем переменных X_t [5,6]:

$$S(X_{t+1} \mid E_{1:t}) = \sum_{X_t} P(X_{t+1}|X_t) S(X_t \mid E_{1:t}). \tag{5}$$

Далее используем метод ВП для определения весов для каждой из выборок в соответствии со свидетельствами $E_{1:t}$:

$$W(X_{t+1} \mid E_{1:t+1}) = P(E_{t+1}|X_{t+1}) S(X_{t+1} \mid E_{1:t}). \tag{6}$$

Для повышения доли согласованных выборок используется процедура повторной генерации выборок в соответствии с распределением $W(X_{t+1} \mid E_{1:t+1})$. Каждая следующая выборка формируется случайным образом. Вероятность того, что будет выбрана конкретная выборка будет пропорциональна весу. Тогда, с учетом этого, распределение по всем выборкам можно записать в следующей окончательном виде [7,8]:

$$\begin{aligned} S(X_{t+1}|E_{1:t+1}) &= N \times W(X_{t+1}|E_{1:t+1}) = \\ &= N \times P(X_{t+1}|E_{1:t+1}). \end{aligned} \tag{7}$$

Алгоритм МЧФ с ВП является оптимальным алгоритмом вероятностного вывода, обладает высокой степенью параллелизма в связи с тем, что генерация выборок может производиться независимо. Однако на ряду с преимуществами, алгоритм обладает недостатком, связанным с тем, что на этапе повторной генерации формирования очередной выборки формируется случайно, тем самым не учитывается предыстория формирования выборок. Для решения данной проблемы в рамках алгоритма МЧФ предлагается использование генетических алгоритмов (ГА). Применительно МЧФ алгоритм ГА может быть использован на этапе повторного формирования выборок с целью снижения разброса выборок относительно истинного значения и как следствие повышения точности определения апостериорного распределения $P(X_{t+1}|E_{t+1})$. Рассмотрим детально структуру генетического алгоритма, а также возможность его адаптации для решения задач вероятностного вывода в СММ на основе филь-

тра МЧФ. Основными этапами генетического алгоритма являются: инициализация, отбор, размножение и мутации. На этапе инициализации выбираем множество выборок, веса которых имеют наибольшие значения. Тогда на этапе отбора доля потомков выборок с высоким значением весов будет преобладающей и будет вытеснять все другие выборки. Тогда каждая следующая популяция будет формироваться путем изменения генотипа предшествующей выборки за счет выполнения процедуры мутации. Введем обобщенную формулировку ГА, предложенную Бэком [9,10] в соответствии с абстрактной моделью Холланда [11]:

$$G(s) =< D_0, n, n_k, \delta, \rho, \Omega, f, t >, \tag{8}$$

где $D_0 = \{d_1, d_2, \ldots, d_n\} \in I^n, I^n = \{0,1\}^{n_k}$ – начальная популяция частиц, полученная на этапе ВП, I^n – множество допустимых значений для X_{t+1}, n – общее число популяция частиц, n_k – размер популяции из которой производим отбор выборок, δ – селекция $I^{n_k} \to I^n$, I^{n_k} – допустимые значения X_{t+1} для популяции n_k, ρ – скрещивание $I^{n_k} \times I^{n_{k+1}} \to I^n$, Ω – мутация $I^n \to I^n$, f – функция соответствия $I^n \to \mathrm{R}$, t – условие останова $I^n \to \{0,1\}$.

Начальная популяция D_0 формируется в соответствии с $\mathrm{P}(\mathrm{E}_{t+1}|\mathrm{X}_{t+1})$ и $\mathrm{P}(X_{t+1}|\mathrm{X}_t)$, процедура оценки весов для каждой S_0^i выполняется в соответствии с алгоритмом ВП. В соответствии с формулой (8) процедура селекции может быть выполнена за счет выборки элементов из множества $D_{t+1} = \{d_1, d_2, \ldots, d_n\}$ в соответствии с равномерным распределением $P(s): I \to \{0,1\}$. Отметим, что в качестве функции соответствия будем использовать весовое распределение $\omega = \mathrm{W}(\mathrm{X}_{t+1}|\mathrm{E}_{1:t+1})$ соответствующее множеству выборок D_{t+1}. В таком случае, наилучший генотип, соответствующей выборке S_{t+1} можно переделить в виде задача поиска экстремума ω:

$$D_{t+1}^i = argmax W(X_{t+1}|E_{1:t+1}). \tag{9}$$

Распределение $\mathrm{P}(D_{t+1}^{\mathrm{i}})$ в общем виде, данное распределение можно выразить через соответствующую функцию соответствия $f(D_{t+1}^{\mathrm{i}}) = \omega(D_{t+1}^{\mathrm{i}})$. Имеем

$$P(D_{t+1}^i) = \frac{\omega(D)}{\sum_{j=1}^{n} \omega(D_{t+1}^j)}. \tag{10}$$

Селекция, представленная выражением (10), предложена Холландом и является адаптивно-пропорциональной (АП). АП также называют методом рулетки. Данный подход обладает существенным недостатком, связанным с тем, что в результате выполнения АП-селекции можно получить отдельные выборки с высоким показателем функции соответствия $\omega(d)$. В таком случае, процедура селекции может быть завершена, так и не достигнув своего оптимального состояния. В связи с этим, в работе будем рассматривать метод Бейкера, позволяющий использовать статистическое распределения для задания $P(d_i)$. Такой подход наиболее оптимальный при решении вероятностных задач на основе МЧФ. Сформулируем селекцию Бейкера (линейное ранжирование) [12] в виде следующего выражения:

$$P(s_i)=\frac{1}{N}\left(a-(a-b)\frac{i-1}{N-1}\right),1\le a\le 2,b=2-a, \tag{11}$$

где i – порядковый номер особи в популяции частиц, N – общий размер популяция выборки, задаваемый на этапе инициализации алгоритма.

Отметим, что параметр a выбирается случайным образом в соответствии с равномерным распределением $\mathrm{U}(1,2)$. При отборе методом линейного ранжирования (ЛР) производим упорядочивание популяции выборок в соответствии с их функцией соответствия, в этом случае данная функция будет эквивалентна весовому распределению ω, соответствующему каждой популяции из выборки D. В таком случае получаем, что распределение $\mathrm{P}(d_{\mathrm{i}})$ будет зависеть только от порядкового номера особи в популяции. Совокупность частиц и соответствующие им веса, упорядоченным согласно весам ω запишем в виде следующего множества:

$$D=\left\{\left\{D_i^1,\omega_i^1\right\},\ldots,\left\{D_i^N,\omega_i^N\right\}\right\},i=1,2,\ldots,n. \tag{12}$$

Для оптимизации линейного ранжирования Бейкера можно воспользоваться разделением выборки частиц на соответствующие схемы с минимальными и максимальными весами. В первые использование механизма схем рассмотрено Холландом. Он предполагал, что для формирования схем можно использовать особи со схожим генотипом. В таком случае схему можно определить в следующем виде

$$\begin{aligned}&H_{S_{t+1}}=\left(\xi=B\,|\,\xi_{i_1}=h_{i_1},\xi_{i_2}=h_{i_2},\ldots,\xi_{i_n}=h_{i_n}\right),\\&i_1<i_2<\ldots<i_n,\end{aligned} \tag{13}$$

где B – множество генотипов, характерное для популяции частиц $\mathrm{S}_{\mathrm{t+1}}^{\mathrm{i}}$, ξ – генотип.

Учитывая тот факт, что один и тот же генотип может быть характерен одновременно для нескольких особей (частиц), можно определить функция приспособленности популяции частиц $D_{\mathrm{t+1}}$ с учетом схемы $\mathrm{H}_{\mathrm{S}_{t+1}}$. Получим [13,18]:

$$f\left(D_{t+1},\Omega_{t+1}\right)=\frac{\sum_{\xi_{t+1}\in H}f\left(\xi_{t+1}^i\right)}{N\left(H_{S_{t+1}},\Omega_{t+1}\right)}, \tag{14}$$

где $\Omega_{\mathrm{t+1}}$ – поколений частиц, соответствующих моменту $t+1$, $N\left(H_{D_{t+1}},\Omega_{t+1}\right)$ – общее число частиц с генотипом ξ в поколении $\Omega_{\mathrm{t+1}}$.

В соответствии с выражением (14) вероятность выживания отдельной популяции частиц будет удовлетворять следующему неравенству:

$$\begin{aligned}&P_l\left(D_{t+1},\Omega_{t+1}\right)\ge 1-\left(1-P_l\left(D_t,\Omega_t\right)\right)\times\\&\times\left(\frac{1-\delta\left(H_{D_{t+1}}\right)}{l-1}\right)\left(\frac{\delta\left(H_{D_{t+1}}\right)}{l-1}\right),\end{aligned} \tag{15}$$

где $\delta(\mathrm{H})$ – длина схемы $\mathrm{H}_{D_{t+1}}$, l – число элементов выборки (частиц).

Для анализа разнообразия популяции выборок можно определить вероятность изменения определенного гена в зависимости от параметров ГА $P\left(\mathrm{N}\left(\mathrm{H}_{D_{t+1}},\Omega_{\mathrm{t+1}}\right)\right)$ в соответствии с

(8). Введем вероятность $\mathrm{P}\left(D_{t+1}^{i},\mathrm{H}_{\mathrm{S}_{t+1}}\right)$, определяющую принадлежность выборки D_{t+1}^{i} схеме $\mathrm{H}_{S_{t+1}}$. Запишем данную вероятность с учетом выражения (14)

$$\begin{aligned} P\left(D_{t+1}^{i},H_{D_{t+1}}\right)&=\frac{N\left(H_{D_{t+1}},\Omega_{t+1}\right)}{N}\times\\ \times\frac{f\left(H_{D_{t+1}},\Omega_{t+1}\right)}{N\left(B,\Omega_{t+1}\right)}&=\\ =\frac{\sum_{\xi_{t+1}\in H_{D_{t+1}}}f\left(\xi_{t+1}^{it}\right)}{\sum_{j=1}^{N}f\left(\xi_{t+1}^{jt}\right)}&, \end{aligned} \tag{16}$$

Тогда выражение для расчета $P\left(\mathrm{N}\left(\mathrm{H}_{\mathrm{S}_{t+1}},\Omega_{\mathrm{t+1}}\right)\right)$ будет иметь следующий вид [15,16]:

$$\begin{aligned} &P\left(N\left(H_{D_{t+1}},\Omega_{t+1}\right)=N\right)=\lambda\left(1-2\theta\right)^{N},\\ &P\left(N\left(H_{D_{t+1}},\Omega_{t+1}\right)=0\right)=\left(1-\lambda\left(1-2\theta\right)\right)^{N},\\ &\lambda=\theta+\left(\frac{1-\delta\left(H_{D_{t+1}}\right)}{l-1}\right), \end{aligned} \tag{17}$$

где $\theta=P\left(D_{t+1}^{i},H_{D_{t+1}}\right)$ – вероятность мутации индивида S_{t+1}^{i} в соответствии со схемой $\mathrm{H}_{\mathrm{S}_{t+1}}$.

В соответствии с выражением (12) можно определить две группы, соответствующие области низких и высоких весов частиц $S_{\mathrm{t+1}}$. Для этого введем критерий разделения. Данный критерий можно получить из нормализации весов ω. Запишем данный критерий в следующем виде:

$$N_{norm}=\frac{1}{\sum_{j=1}^{N}\left(\omega_{j}^{i}\right)^{2}} \tag{18}$$

Тогда выборку S_{t+1}, соответствующую моменты времени $t+1$ можно записать в виде совокупности двух групп, соответствующих области высоких и низких весов, разделяемых в соответствии с $\mathrm{N}_{\mathrm{norm}}$. Получим [17,18]:

$$\begin{aligned} &H_{t+1}=\begin{cases} H_{t+1}^{'}=\left\{\left\{\xi_{i_1},\omega_{i_1}\right\},\ldots,\left\{\xi_{i_k},\omega_{i_k}\right\}\right\}\\ H_{t+1}^{''}=\left\{\left\{\xi_{i_{k+1}},\omega_{i_{k+1}}\right\},\ldots,\left\{\xi_{i_N},\omega_{i_N}\right\}\right\} \end{cases},\\ &k\le N_{norm}<k+1, \end{aligned} \tag{19}$$

где $\mathrm{H}_{\mathrm{t+1}}^{'}$ и $\mathrm{H}_{\mathrm{t+1}}^{''}$ – схемы выборки, соответствующие области наибольших и наименьшим весов выборок S_{t+1}.

Далее перейдем к операции скрещивания популяции частиц для формирования оптимальной выборки. Основной задачей скрещивания является повышения разнообразия популяции частиц, что позволяет получить предельно неповторяющийся набор генотипов частиц в рамках фиксиро-

ванного момента времени $t+1$. Для этого воспользуемся равномерным кроссинговером (РК). В классической интерпретации производится равновероятностное копирование генов от нескольких родителей к потомку. Для усиления равномерного кроссинговера, вместо равновероятностного выбора родительских вершин будет использовать метод взвешивания с учетом правдоподобия, а именно для каждой новой генерации оцениваются веса возможных родителей. Тогда процесс скрещивания можно представить в виде следующих уравнений:

$$D_{t+1}^{i,j} = \begin{cases} D_{t+1}^{i} = \alpha D_t^i + (1-\beta) D_t^j \\ D_{t+1}^{j} = \beta D_t^i + (1-\alpha) D_t^i \end{cases},$$

$$\alpha = \frac{W\left(D_t^i\right)}{W\left(D_t^i\right) + W\left(D_t^j\right)}, \tag{20}$$

$$\beta = \frac{W\left(D_t^j\right)}{W\left(D_t^i\right) + W\left(D_t^j\right)},$$

где $D_{t+1}^{i,j}$ – новая популяция частиц, образованная путем скрещивания родителей D_t^i и D_{t+1}^i из выборки S_t, α, β – весовые коэффициенты, устанавливающие соответствие между весами родителей $W\left(D_t^i\right)$ и $W\left(D_t^j\right)$.

Процедура мутации может быть реализована на основе распределения Гаусса. Каждая мутирующий представитель выборки $D_{t+1} = \left(d_{t+1}^1, d_{t+1}^2, \ldots, d_{t+1}^n\right)$ будет формироваться в соответствии со следующим выражением:

$$\begin{aligned} d_{t+1}^{'i} &= M\left(d_t^i\right), \\ M\left(d_t^i\right) &= d_t^i + N\left(0, \sigma_i\right), \end{aligned} \tag{21}$$

где $N\left(0, \sigma_i\right)$ – гауссовское распределение, d_{t+1}^i – популяция выборки, полученная на шаге $t+1$, на основе которой производится мутация, $M\left(d_t^i\right)$ – процедура мутации для популяции d_t^i.

Приведем укрупненную схему алгоритма фильтрации частиц, используемого в исследовании:

Шаг 1. На начальном срезе из распределения $P\left(X_0\right)$ одновременно генерируется N выборок.

Шаг 2. Вводится множества свидетельств для всех срезов сети $E_1, E_2, ..., E_T$.

Шаг 3. Выполняется перехода от временного среза t к временному срезу $t+1$. Через модель перехода $P\left(X_{t+1} \middle| X_t\right)$ осуществляется обновление множества выборок: $N\left(X_{t+1} \middle| E_{1:t}\right) = \sum_{X_t} P\left(X_{t+1} \middle| X_t\right) N\left(X_t \middle| E_{1:t}\right)$, $N\left(X_t \middle| E_{1:t}\right)$ – количество выборок для состояния X_t после получения свидетельств $E_{1:t}$) выборки взвешиваются с учетом правдоподобия по отношению к новым свидетельствам E_{t+1}, им присваивается вес $P\left(E_{t+1} \middle| X_{t+1}\right)$.

Шаг 4. Вычисляется суммарный вес выборок в состоянии X_{t+1} после получения свидетельств E_{t+1}: $w\left(X_{t+1}\middle|E_{1:t+1}\right)=P\left(E_{t+1}\middle|X_{t+1}\right)P\left(X_{t+1}\middle|E_{1:t}\right)$ отбрасываются выборки с малым весом формируются новые N выборок.

Шаг 5. Формируется новая популяция, состоявшая из N выборок с использование генетического алгоритма, каждая выборка тиражируется пропорционально весу $w\left(X_{t+1}\middle|E_{1:t+1}\right)$. Каждая выборка образуется путем скрещивания родителей D_{t+1}^{i} и D_{t+1}^{i+1} из выборки S_{t+1}, полученной на предыдущем шаге. Процедура мутации выполняется в соответствии с выражением (21).

Оценка эффективности применения генетических алгоритмов для решения вероятностных задач тестирования

Оптимизация фильтра МЧФ на основе ГА является важным алгоритмом оптимизации стохастических процедур вероятностного вывода. Для сравнения алгоритмов МЧФ и МЧФ с ГА проведения эксперимента разработаны распределенные алгоритмы, использование которых предусмотрено в рамках параллельной платформы Spark [19], развернутой в облачной среде Yandex Cloud из 6 узлов со следующей аппаратной конфигурацией: 2 процессора Intel Xeon-Platinum 2.5 ГГц 16 ядер, 128 ГБ ОЗУ, жесткий диск 10 ТБ, оптический канал связи 25 Гб/с. По результатам исследования нами решена задача повышения качества формирования выборок на основе МЧФ фильтра. Ее решение заключается в достижении требуемой точности алгоритма за счет применения ГА, что в свою очередь позволяет повысить долю выборок, согласованных со свидетельствами. Применение генетических алгоритмов доказывает правильность выбора исследования, а также состоятельность предложенного подхода.

В рамках практической проверки предложенного метода оптимизации МЧФ на основе ГА произведено его сравнение с классическим МЧФ за счет оценки согласованности выборок на этапе повторной генерации выборок в процессе вероятностного вывода СММ процесса тестирования веб-приложений (441 узел, 195 ребер, 72079 параметров). Каждый узел СММ отвечает зав формирования определенного набора тестовых данных для поиска программных ошибок, связанных с аутентификацией пользователей за счет использования SQL-инъекции. Данная ошибка представляет собой набор SQL-команд с разделителями, позволяющая обходит фильтры безопасности веб-приложений и реализующая возможность получения пользовательские данных, используемые для аутентификации и авторизации. На сегодняшний день можно выделить следующие типы SQL-инъекций, которые могут быть использованы для получения пользовательских данных из веб-приложений, взаимодействующих с системами управления базами данных: Union (Объединенная), Boolean Blind (Логическая.), Time Blind (Временная), Error Blind (На основе ошибок), Stacked Queries (Разделенная) и Out of Band (Внешняя). В таблице 1 приведем основные показатели согласованности $\psi_{S_{t+1}}$ выборок S_{t+1} и свидетельств E_{t+1} в процессе тестирования веб-приложений на предмет наличия ошибок типа SQLинъекции в зависимости от общего числа свидетельств во всех временных срезах N_e для данной сети с общим объемом выборок $N_s=1000000$. Отметим, что второй и последующие срезы СММ характеризуются применением межсетевого экрана ModSecurity для блокирования типичных ошибок и оптимизацию формирования выборок для выявления аномальных ошибок.

Таблица 1 – Сравнение степени согласованности выборок

№	Алгоритм	Ne=200	Ne=500	Ne=1000	Ne=5000
1	Алгоритм МЧФ	10%	15%	18%	25%
2	Алгоритм МЧФ с ГА	3%	3%	3%	3%

Из таблицы 1 получаем важный практический результат, заключающийся в том, что при использовании алгоритма ГА совместно с МЧФ наблюдаем фиксированную степень согла-

сованности выборок вне зависимости от числа переменных свидетельств N_e. Следовательно каждая результирующая популяция, полученная по итогам выполнения алгоритма ГА, будет иметь наибольшее значение функции приспособленности ω, что ведет к повышению точности апостериорного распределения вероятностей $P\left(\mathrm{X}_{t+1}|\mathrm{E}_{t+1}\right)$. Другая особенность заключается в том, что при использовании ГА можно ограничить общее число первоначально формируемых выборок N, при этом можно повысить число шагов мутаций на этапе повторного взвешивания. Такой подход позволяет настроить точность алгоритма за минимальное число шагов алгоритма МЧФ. Отметим, что применение алгоритма ГА без МЧФ к СММ не имеет смысла, так как в процессе выполнения генетического алгоритма нет необходимой информации относительно свидетельств, а также вероятностных связей между тиражируемыми между срезами переменными. С помощью ГА можно оптимизировать существующую выборку, полученную по результатам выполнения МЧФ-фильтрации. Следовательно, для повышения эффективности МЧФ можно использовать различные подходы к его распараллеливанию, в таком случае формирование выборок S_{t+1} будет выполняться независимо друг от друга. В таблице 2 приведем сравнительные характеристики двух программных реализаций алгоритмов МЧФ и МЧФ с ГА.

Таблица 2 – Сравнение производительности программных реализаций алгоритмов

№	Алгоритм	Однопоточный режим	Параллельный режим на 1 узле	Параллельный режим на 6 узлах
1	МЧФ	844,28878670758 сек.	648,15711330071 сек.	305,911736891564 сек.
2	МЧФ с ГА	851,96107700558 сек.	656,059819253189 сек.	321,565701030924 сек.

Из анализа таблицы 2 следует, что применение ГА в незначительной степени сказывается на производительности алгоритма. Однако при использовании МЧФ с ГА количество выборок необходимых для достижения требуемого уровня согласованности $\psi_{S_{t+1}}$ может быть существенно сокращено за счет повышения доли популяций, согласованных со свидетельствами E_{t+1}.

Заключение и выводы

Решение задач оптимизации, существующих алгоритмов вероятностного вывода является актуальным направлением исследования. В первую очередь это связано с повышением сложности вероятностных моделей, повышением числа скрытых переменных, а также роста потока свидетельств на каждом из временных срезов. В работе рассмотрено применение предложенных подходов к скрытым марковским моделям, однако данные алгоритмы могут быть адаптированы для реализации процедуры логического вывода, как в статических, так и динамических вероятностных моделях. Использование ГА в процессе МЧФ-фильтрации позволяет решить задачу качества формирования выборок, взамен использования случайной выборки в процессе повторной генерации используется генетический алгоритмы. Такой подход позволяет повысить точность апостериорного распределения с условий роста переменных-свидетельств E_{t+1}, а также в полной мере использовать алгоритм взвешивая с учетом правдоподобия для формирования весовых распределений для первичной популяции выборок $D_0 = \left(d_1, d_2, \ldots, d_n\right)$. Применение модели схем Холланда позволяет разграничить области выборок с разными генотипами в соответствии с распределением $\mathrm{W}\left(\mathrm{X}_{t+1}|\mathrm{E}_{1:t+1}\right)$. Это достигается за счет того, что в рамках ГА весовое распределение выбирается в качестве функции соответствия, устанавливающей соответствия весов и каждой популяции частиц, входящей в состав выборки. Отметим, что предложенный алгоритм обладает схожей с классическим МЧФ производительности, однако позволят повысить область согласования выбо-

рок в соответствии с потоком свидетельств $\mathrm{E}_{1:t+k} = (e_1, e_2, \ldots, e_n)$, поступающих вплоть до момента времени $t + k$. Разработанный алгоритм является достаточно хорошо масштабируемым и его можно распараллелить. В этом случае процедура формирования выборок S_{t+1} может выполняться независимо. Процедуру распараллеливания ГА можно реализовать на этапе скрещивания и мутации, так как отбор родителей особи каждой следующей популяции выбирается в соответствии с весами потомков, которые уже заранее известны. Практические результаты анализа эффективности предложенного алгоритма позволяют прийти к выводу, что незначительно снижение производительности за счет использования в процессе выполнения ГА операций селекции, скрещивания и мутации по сравнению с классическим алгоритмом МЧФ позволяют повысить точность апостериорного распределении и степень согласования выборок в соответствии со свидетельствами, вне зависимости от общего их числа. Такой подход позволяет использовать разработанный алгоритм при решении задач логического вывода в различных динамических вероятностных моделях с неограниченным числом временных состояний. Все это доказывает обоснованность и практическую значимость проведенных исследований.

Литература

1. Костенко, В. А. Генетически алгоритм с самообучением / В. А. Костенко, А. В. Фролов. – Текст : непосредственный // Известия РАН. Теория и системы управления. – 2015. – № 4. – С. 24–38.

2. Moral, P. D. On the concentration Properties of interacting particle processes / P. D. Moral, P. Hu, L. Wu // Machine learning. – 2010. – Vol. 3, № 4, – P. 225–389.

3. Moral, P. D. Particle filter: An introduction with application / P. D. Moral, A. Doucet // ESAIM. – 2014. – Vol. 14. – P. 1–46.

4. Рыбаков, К. А. Непрерывные фильтры частиц и их реализация в реальном масштабе времени / К. А. Рыбаков, А. А. Ющенко. – Текст : непосредственный // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2018. – № 3. – С. 56–64.

5. Евстигнеев, М. И. Локализация мобильного робота с фильтром частиц при обнаружении и сегментации объектов / М. И. Евстигнеев, Ю. В. Литвинов, В. В. Мазулина. – Текст : непосредственный // Научно-технический вестник информационных технологий, механики и оптики. – 2019. – Т. 19, № 4. – С. 622–629.

6. Волков, В. А. Численное решение задач нелинейной фильтрации на основе алгоритмов фильтра частиц / В. А. Волков, И. А. Кудрявцев. – Текст : непосредственный // Труды МАИ. – 2016. – № 89. – С. 1–21.

7. Тулупьев, А. Л. Апостериорные оценки вероятностей в алгебраических байесовских сетях / А. Л. Тулупьев. – Текст : непосредственный // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика. Информатика. Процессы управления. – 2012. – № 2. – С. 51–59.

8. Yin, S. Intelligent particle filter and its application to fault detection of nonlinear system / S. Yin, X. Zhu // IEEE Transactions on Industrial Electronics. – 2015. – Vol. 62, № 5. – P. 3852–3861.

9. Russel, S. Artificial intelligence a modern approach, 4 edition / S. Russel, P. Norvig. – Hoboken : Pearson, 2020. – 1023 p.

10. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter / M. K. Pitt, R. S. Silva, P. Giordani, R. Kohn // Journal of Econometrics. – 2012. – Vol. 171, № 2. – P. 134–151.

11. Golightly, A. Bayesian parameter inference for stochastic biochemical / A. Golightly, D. J. Wilkinson // Interface Focus. – 2011. – Vol. 1, № 6. – P. 807–820.

12. Bunch, P. Improved particle approximations to the joint smoothing distribution using Markov chain Monte Carlo / P. Bunch, S. Godsill // IEEE Transactions on Signal Processing. – 2013. – Vol. 61, № 4. – P. 956–953.

13. Azarnova, T. V. Advanced hybrid stochastic dynamic Bayesian network inference algorithm development in the context of the web applications test execution / T. V. Azarnova, P. V. Polukhin // Journal of Physics: Conf. Series: Materials Science and Engineering. – 2019. – Vol. 537. – P. 052028.

14. Белых, М. А. Схема работы выбора эволюционного алгоритма интеллектуальной системы / М. А. Белых, А. В. Барабанов. – Текст : непосредственный // Информационные технологии моделирования и управления. – 2021. – Т. 128, № 128. – С. 114–117.

15. Галуа, Д. В. Об алгоритме возмущения полудискретной схемы для эволюционных уравнений и оценки погрешности приближенного решения с помощью полугрупп / Д. В. Галуа, Д. Л. Рогаева. – Текст : непосредственный // Журнал вычислительной математики и математической физики. – 2016. – Т. 59, № 7. – С. 1299–1322.

16. Структура интеллектуальной системы поддержки эволюционных алгоритмов / М. А. Белых, В. Ф. Барабанов, С. Л. Подвальный, А. К. Донских. – Текст : непосредственный // Вестник Воронежского государственного технического университета. – 2021. – Т. 17, № 3. – С. 7–13.

17. Лотов, А. В. Простая эффективная гибридизация классической глобальной оптимизации и генетических алгоритмов многокритериальной оптимизации / А. В. Лотов, А. И. Рябиков. – Текст : непосредственный // Журнал вычислительной математики и математической физики. – 2019. – Т. 59, № 10. – С. 1666–1680.

18. Денисова, Л. А. Проектирование систем управления на основе многокритериальной оптимизации с использованием генетических алгоритмов / Л. А. Денисова, В. А. Мещерякова. – Текст : непосредственный // Автоматизация в промышленности. – 2015. – № 10. – С. 18–24.

19. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing / M. Zaharia, M. Chowdhury, T. Das [et al.] // NSDI. – 2012. – P. 1–15.